THE MAXIMUM ENTROPY ON THE MEAN METHOD, CORRELATIONS AND IMPLEMENTATION ISSUES

C. HEINRICH, J.–F. BERCHER AND G. DEMOMENT
Laboratoire des signaux et systèmes
Plateau de Moulon
91192 Gif–sur–Yvette Cedex, France †

Abstract. In this paper, we address the problem of solving linear inverse problems using the Maximum Entropy on the Mean Method (MEMM). The observation equation $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ links the observations $\boldsymbol{z}$ and the supposed known degradation matrix $\boldsymbol{H}$ to the object $\boldsymbol{x}$ and the noise $\boldsymbol{n}$ which have to be estimated. Given a reference measure ν defined jointly on $\boldsymbol{x}$ and $\boldsymbol{n}$, the MEMM consists in selecting the distribution which is closest to ν according to the Kullback distance and whose mean satisfies the observation equation. The MEMM estimator $\hat{\boldsymbol{x}}$ is the mean of the selected distribution. This amounts to selecting $\hat{\boldsymbol{x}}$ as the minimizer of a convex cost function defined on $\boldsymbol{x}$ and $\boldsymbol{n}$ [1].

Up to now, this method yielded only separable objective functions. We present here an extension of the method, allowing penalization of linear functions of $\boldsymbol{x}$ (such as differences between pixels for example) in the MEMM formalism. We also discuss algorithmic issues: the method amounts to minimizing a convex criterion and thus admits a dual formulation. We will take advantage of both formulations to reduce computational effort, which hadn't been achieved to this day since the problem had only been solved in the dual variables. We report two simulation examples, a deconvolution problem in spectroscopy and a Fourier synthesis example.

Key words: maximum entropy on the mean, maximum entropy, convex cost functions, duality, Legendre–Fenchel transform.

1. Introduction

We consider here linear inverse problems, whose observation model satisfies $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$. We want to estimate the object $\boldsymbol{x}$ and the noise $\boldsymbol{n}$, knowing the observations $\boldsymbol{z}$ and the degradation matrix $\boldsymbol{H}$. Matrix $\boldsymbol{H}$ will typically be a convolution or a Fourier matrix and vector $\boldsymbol{x}$ a signal or an image. A typical resolution framework

†Email: heinrich@lss.supelec.fr

for this type of ill–posed problems is the minimization of a composite criterion

$$\mathcal{F}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}),$$

yielding $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{F}(\boldsymbol{x})$. This objective function comprises a regularization term $\mathcal{F}_{\boldsymbol{x}}$ and a term $\mathcal{F}_{\boldsymbol{n}}$ penalizing the residual noise $\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}$. Typical choices [2] are $\mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) = \lambda_1 \sum_i \left|\frac{x_i}{\Delta_1}\right|^p + \lambda_2 \sum_j \left|\frac{x_{j+1}-x_j}{\Delta_2}\right|^q$ and $\mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}) = \| \boldsymbol{z} - \boldsymbol{H}\boldsymbol{x} \|^2$.

Convexity of the cost function is a very important issue. Nonconvex cost functions are cumbersome to minimize and the estimator may not be a continuous function of the observations, continuity being a desirable property. Nevertheless, nonconvex cost functions may give better results than convex cost functions, especially regarding edge restoration. On the other hand, convex cost functions can be minimized using classical descent algorithms, such as gradient or conjugate gradient. Those functions guarantee continuity of the estimator. They are minimized at a reasonable computational cost.

Remembering Bayes' rule $p(\boldsymbol{x}|\boldsymbol{z}) \propto p(\boldsymbol{z}|\boldsymbol{x})\, p(\boldsymbol{x})$ linking the *a posteriori* distribution $p(\boldsymbol{x}|\boldsymbol{z})$ to the likelihood $p(\boldsymbol{z}|\boldsymbol{x})$ and the *a priori* distribution $p(\boldsymbol{x})$, $\hat{\boldsymbol{x}}$ can always be interpreted as a maximum *a posteriori* (MAP) estimator, providing the relations

$$\begin{aligned} p(\boldsymbol{x}) &\propto \exp\left(-\mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x})\right), \\ p(\boldsymbol{z}|\boldsymbol{x}) &\propto \exp\left(-\mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})\right) \end{aligned}$$

are satisfied.

We deal here with an alternative framework, the maximum entropy on the mean method, which allows to both construct and interpret convex cost functions. This class of convex cost functions comprises a lot of well known criteria, such as Shannon and Burg entropies and L_2 norm for example [3,1]. The method is closely linked to statistical mechanics. It was introduced by Navaza [4,5] to solve a inverse problem in crystallography. It has been further investigated by Dacunha–Castelle, Gamboa and Gassiat [3,6] from a mathematical point of view. Application of the method to inverse problems has been studied by Bercher, Le Besnerais and Demoment [7,1,8].

Given a reference measure ν defined on the objects to be estimated, this method consists in selecting the distribution $\hat{p}$ minimizing the Kullback distance to ν among the distributions whose mean satisfies a given constraint. The MEM estimator is the mean of the selected distribution. The constraint is here an observation equation which will be assumed linear. Dealing with nonlinear problems is possible in the MEM formalism, but the optimization problem may no longer be convex.

We will assume that the constraints are qualified for all inverse problems considered here, which means that there exists a least one solution to any problem.

2. The maximum entropy on the mean method

To give a first sketch of the method, we will consider the problem of selecting an estimator $\hat{\boldsymbol{x}}$ among two candidates $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$. We will show that the MEM procedure amounts to selecting a distribution among an exponential family, and that deriving

the MEM estimator from this distribution is equivalent to minimizing a convex cost function defined on the object to be estimated.

Let

$$\mathcal{K}(p,\nu) = \begin{cases} \int \log \dfrac{\mathrm{d}p}{\mathrm{d}\nu}\, \mathrm{d}p & \text{if } p \ll \nu, \\ +\infty & \text{otherwise,} \end{cases}$$

be the Kullback distance between the distribution p and the reference measure ν. Following the MEM principle, we derive:

$$\begin{array}{lcl} \hat{p} & = & \arg\min_{p\in\mathcal{C}}\ \mathcal{K}(p,\nu)\,, \\ \mathcal{C} & = & \mathcal{C}_{\boldsymbol{x}_1} \cup \mathcal{C}_{\boldsymbol{x}_2}\,, \\ \mathcal{C}_{\boldsymbol{x}_i} & = & \left\{ p \in \mathcal{M}_1^+ \,|E_p[\boldsymbol{u}] = \boldsymbol{x}_i \right\}, \\ \hat{\boldsymbol{x}} & = & E_{\hat{p}}[\boldsymbol{u}]\,, \end{array}$$

where $\mathcal{M}_1^+$ is the set of the probability measures.

Let $p_i = \arg\min_{p\in\mathcal{C}_{\boldsymbol{x}_i}}\ \mathcal{K}(p,\nu)$. We have $\hat{p} = \arg\min_{p\in\{p_1,p_2\}} \mathcal{K}(p,\nu)$, since

$$\forall p \in \mathcal{C} \qquad \mathcal{K}(p,\nu) \geq \min\left(\ \mathcal{K}(p_1,\nu)\,, \mathcal{K}(p_2,\nu)\ \right).$$

Selecting $\hat{\boldsymbol{x}}$ in $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ thus amounts to comparing $\mathcal{F}(\boldsymbol{x}_1) \triangleq \mathcal{K}(p_1,\nu)$ to $\mathcal{F}(\boldsymbol{x}_2) \triangleq \mathcal{K}(p_2,\nu)$.

This can be extended to the choice of an estimator among any family $\mathcal{E}$ (whether convex or not) of candidates.

We still have $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathcal{E}}\ \mathcal{F}(\boldsymbol{x})$, where $\mathcal{F}(\boldsymbol{x}) = \min_{\{p\,|E_p[\boldsymbol{u}]\in\mathcal{E}\}}\ \mathcal{K}(p,\nu)$. Convexity of $\mathcal{F}(.)$ derives straightforwardly from convexity of $\mathcal{K}(.,\nu)$.

In the large deviations theory, $\mathcal{F}(.)$ is said to be a level-1 entropy, whereas $\mathcal{K}(.,\nu)$ is the level-2 entropy (see [9] for developments on the large deviations theory and [10] for the links of the MEMM with statistical mechanics).

Let us now focus on distribution p_1 defined previously, so as to show that p_1 as well as $\hat{p}$ belong to an exponential family. According to its definition, distribution p_1 minimizes the Kullback distance among all distributions of mean $\boldsymbol{x}_1$. This convex constrained minimization problem is solved using the Lagrange multipliers method, which requires to find the stationary points of the Lagrangian

$$L = \mathcal{K}(p,\nu) + \boldsymbol{\mu}^t\left(\boldsymbol{x}_1 - E_p[\boldsymbol{u}]\right) + \mu_0\left(1 - E_p[1]\right)$$

associated to the problem. Vector $\boldsymbol{\mu}$ and scalar μ_0 are the Lagrange multipliers associated to the constraints. The Lagrangian will have to be minimized in variable p. We have:

$$L = \int \frac{\mathrm{d}p}{\mathrm{d}\nu}\left(\log\frac{\mathrm{d}p}{\mathrm{d}\nu} - \boldsymbol{\mu}^t\boldsymbol{u} - \mu_0\right)\,\mathrm{d}\nu + \boldsymbol{\mu}^t\boldsymbol{x}_1 + \mu_0\,.$$

Moreover, from

$$\forall y \in \mathrm{R}^{+*},\ \forall \alpha \in \mathrm{R} \qquad y\left(\log y - \alpha\right) \geq -\exp\left(\alpha - 1\right),$$

we derive

$$\frac{\mathrm{d}p}{\mathrm{d}\nu}\left(\log\frac{\mathrm{d}p}{\mathrm{d}\nu} - \boldsymbol{\mu}^t\boldsymbol{u} - \mu_0\right) \geq -\exp\left(\boldsymbol{\mu}^t\boldsymbol{u} + \mu_0 - 1\right).$$

Hence

$$\forall p \in \mathcal{M}^+ \qquad L \geq \mu_0 + \boldsymbol{\mu}^t \boldsymbol{x}_1 - \int \exp\left(\boldsymbol{\mu}^t \boldsymbol{u} + \mu_0 - 1\right) \, \mathrm{d}\nu,$$

with equality if and only if

$$\frac{\mathrm{d}p\left(\boldsymbol{u}\right)}{\mathrm{d}\nu\left(\boldsymbol{u}\right)} = \exp\left(\boldsymbol{\mu}^t \boldsymbol{u} + \mu_0 - 1\right).$$

Let us define

$$\mathcal{F}^*\left(\boldsymbol{\mu}\right) \stackrel{\Delta}{=} 1 - \mu_0 = \log \int \exp\left(\boldsymbol{\mu}^t \boldsymbol{u}\right) \mathrm{d}\nu\left(\boldsymbol{u}\right),$$

assuring normalization of

$$\mathrm{d}p\left(\boldsymbol{u}\right) = \exp\left(\boldsymbol{\mu}^t \boldsymbol{u} + \mu_0 - 1\right) \mathrm{d}\nu\left(\boldsymbol{u}\right) = \exp\left(\boldsymbol{\mu}^t \boldsymbol{u} - \mathcal{F}^*\left(\boldsymbol{\mu}\right)\right) \mathrm{d}\nu\left(\boldsymbol{u}\right). \tag{1}$$

If we combine expressions of the Lagrangian L and of the exponential family (1), the Lagrangian reduces to $D\left(\boldsymbol{\mu}\right) \stackrel{\Delta}{=} \boldsymbol{\mu}^t \boldsymbol{x}_1 - \mathcal{F}^*\left(\boldsymbol{\mu}\right)$. Providing this concave function can be maximized, we define $\hat{\boldsymbol{\mu}}_1 = \arg\max_{\boldsymbol{\mu}} D\left(\boldsymbol{\mu}\right)$. This defines the distribution $\hat{p}_1$ which is a stationary point of the Lagrangian L. By the way, we also established that $\mathcal{K}\left(\hat{p}_1, \nu\right) = \mathcal{F}\left(\boldsymbol{x}_1\right) = D\left(\hat{\boldsymbol{\mu}}_1\right)$ (dual attainment relation).

We derive at last equation $\mathcal{F}\left(\boldsymbol{x}\right) = \max_{\boldsymbol{\mu}}\left(\boldsymbol{\mu}^t \boldsymbol{x} - \mathcal{F}^*\left(\boldsymbol{\mu}\right)\right)$, linking $\mathcal{F}$ and $\mathcal{F}^*$ as convex-conjugates or Legendre–Fenchel transforms of each other. Developments on the Legendre–Fenchel transform may be found in [9]. $\mathcal{F}^*$ is said to be the log-partition function of ν, and $\mathcal{F}$ is said to be the Cramér transform of ν. This uniquely defines $\mathcal{F}$ if one is given a reference measure ν.

It must also be emphasized that the method wasn't exposed this way up to day. The former presentation (see [10] for example) restricted the application framework of the method, since it supposed that the observation equation was necessarily linear. This new presentation allows to deal with nonlinear problems and also allows to select an object among a discrete set of candidates. Of course, those are merely extensions enlightened by the presentation. These potentialities already existed before, but might have been hidden by the presentation.

3. The MEMM applied to linear inverse problems

In this section, we apply the MEMM reviewed in the previous section to the resolution of linear inverse problems. We will detail two equivalent formulations of the MEMM. We review additional developments needed to circumvent analytical difficulties otherwise seriously limiting the method.

We use notations previously established. The unknowns to be estimated are the object $\boldsymbol{x}$ and the noise $\boldsymbol{n}$. The reference measure will have to be defined on $\boldsymbol{x}$ and $\boldsymbol{n}$, and chosen such that $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{n}})$ may be considered as the mean of a distribution near ν according to the Kullback distance. In particular, the selection procedure of $\hat{p}$ implies that $\hat{p}$ has the same support than ν. Since $\hat{\boldsymbol{x}}$ is the mean of $\hat{p}$, $\hat{\boldsymbol{x}}$ belongs to the convex hull of the support of ν. We take advantage of this property to constrain $\hat{\boldsymbol{x}}$ to belong to a given set. A reasonable choice is a separable reference measure $\mathrm{d}\nu\left(\boldsymbol{u}, \boldsymbol{v}\right) = \mathrm{d}\nu_{\boldsymbol{x}}\left(\boldsymbol{u}\right) \, \mathrm{d}\nu_{\boldsymbol{n}}\left(\boldsymbol{v}\right)$.

3.1. PRIMAL FORMULATION

We define here $\mathrm{d}\nu(\boldsymbol{u},\boldsymbol{v}) \triangleq \mathrm{d}\nu_{\boldsymbol{x}}(\boldsymbol{u})\ \mathrm{d}\nu_{\boldsymbol{n}}(\boldsymbol{v})$ and $\boldsymbol{\mu} \triangleq [\boldsymbol{\mu}_{\boldsymbol{x}}^t, \boldsymbol{\mu}_{\boldsymbol{n}}^t]^t$, $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\boldsymbol{u}$ having the same size as $\boldsymbol{x}$, $\boldsymbol{\mu}_{\boldsymbol{n}}$ and $\boldsymbol{v}$ having the same size as $\boldsymbol{n}$. Following the reasoning exposed in the previous section, the log-partition function $\mathcal{F}^*$ of the reference measure ν is given by:

$$\mathcal{F}^*(\boldsymbol{\mu}) = \log \int \exp\left(\boldsymbol{\mu}_{\boldsymbol{x}}^t \boldsymbol{u} + \boldsymbol{\mu}_{\boldsymbol{n}}^t \boldsymbol{v}\right)\ \mathrm{d}\nu(\boldsymbol{u},\boldsymbol{v}).$$

We derive:

$$\begin{aligned}\mathcal{F}^*(\boldsymbol{\mu}) &= \log \int \exp\left(\boldsymbol{\mu}_{\boldsymbol{x}}^t \boldsymbol{u}\right)\ \mathrm{d}\nu_{\boldsymbol{x}}(\boldsymbol{u}) + \log \int \exp\left(\boldsymbol{\mu}_{\boldsymbol{n}}^t \boldsymbol{v}\right)\ \mathrm{d}\nu_{\boldsymbol{n}}(\boldsymbol{v}), \\ &\triangleq \mathcal{F}_{\boldsymbol{x}}^*(\boldsymbol{\mu}_{\boldsymbol{x}}) + \mathcal{F}_{\boldsymbol{n}}^*(\boldsymbol{\mu}_{\boldsymbol{n}}).\end{aligned}$$

Hence we have:

$$\begin{aligned}\mathcal{F}(\boldsymbol{x},\boldsymbol{n}) &= \max_{\boldsymbol{\mu}_{\boldsymbol{x}},\boldsymbol{\mu}_{\boldsymbol{n}}} \left(\boldsymbol{\mu}_{\boldsymbol{x}}^t \boldsymbol{x} + \boldsymbol{\mu}_{\boldsymbol{n}}^t \boldsymbol{n} - \mathcal{F}_{\boldsymbol{x}}^*(\boldsymbol{\mu}_{\boldsymbol{x}}) - \mathcal{F}_{\boldsymbol{n}}^*(\boldsymbol{\mu}_{\boldsymbol{n}})\right), \\ &= \max_{\boldsymbol{\mu}_{\boldsymbol{x}}} \left(\boldsymbol{\mu}_{\boldsymbol{x}}^t \boldsymbol{x} - \mathcal{F}_{\boldsymbol{x}}^*(\boldsymbol{\mu}_{\boldsymbol{x}})\right) + \max_{\boldsymbol{\mu}_{\boldsymbol{n}}} \left(\boldsymbol{\mu}_{\boldsymbol{n}}^t \boldsymbol{n} - \mathcal{F}_{\boldsymbol{n}}^*(\boldsymbol{\mu}_{\boldsymbol{n}})\right), \\ &\triangleq \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{n}).\end{aligned}$$

This shows that the cost function $\mathcal{F}$ defined on $\boldsymbol{x}$ and $\boldsymbol{n}$ is separable if the reference measure is separable. The MEM estimator is selected by minimizing $\mathcal{F}$ over all objects satisfying the observation equation, *i.e.* over all objects satisfying $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$. This yields $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})$, which has the classical form of a regularized cost function. This is said to be the primal formulation of the MEM, and $\mathcal{F}$ is called the primal cost function. We will now derive the dual formulation.

3.2. DUAL FORMULATION

To the minimization of a convex function (primal formulation) can be associated the maximization of a concave function (dual formulation). Both formulations yield the same solution. This relies on the Legendre–Fenchel duality theory [9]. We have:

$$\mathcal{F}_{\boldsymbol{x}}(\hat{\boldsymbol{x}}) + \mathcal{F}_{\boldsymbol{n}}(\hat{\boldsymbol{n}}) = \max_{\boldsymbol{\mu}_{\boldsymbol{x}},\boldsymbol{\mu}_{\boldsymbol{n}}} \left(\boldsymbol{\mu}_{\boldsymbol{x}}^t \hat{\boldsymbol{x}} + \boldsymbol{\mu}_{\boldsymbol{n}}^t \hat{\boldsymbol{n}} - \mathcal{F}_{\boldsymbol{x}}^*(\boldsymbol{\mu}_{\boldsymbol{x}}) - \mathcal{F}_{\boldsymbol{n}}^*(\boldsymbol{\mu}_{\boldsymbol{n}})\right).$$

From $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})$, we derive:

$$\left.\frac{\partial \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} - \boldsymbol{H}^t \left.\frac{\partial \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{n})}{\partial \boldsymbol{n}}\right|_{\boldsymbol{n}=\boldsymbol{z}-\boldsymbol{H}\hat{\boldsymbol{x}}} = 0.$$

Hence $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} - \boldsymbol{H}^t \hat{\boldsymbol{\mu}}_{\boldsymbol{n}} = 0$ since, as a result of Fenchel duality,

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \left.\frac{\partial \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \quad \text{and} \quad \hat{\boldsymbol{\mu}}_{\boldsymbol{n}} = \left.\frac{\partial \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{n})}{\partial \boldsymbol{n}}\right|_{\boldsymbol{n}=\hat{\boldsymbol{n}}}.$$

We have at last:

$$\mathcal{F}_{\boldsymbol{x}}(\hat{\boldsymbol{x}}) + \mathcal{F}_{\boldsymbol{n}}(\hat{\boldsymbol{n}}) = \max_{\boldsymbol{\mu_n}} \left(\boldsymbol{\mu}_{\boldsymbol{n}}^t \boldsymbol{z} - \mathcal{F}_{\boldsymbol{x}}^* \left(\boldsymbol{H}^t \boldsymbol{\mu}_{\boldsymbol{n}} \right) - \mathcal{F}_{\boldsymbol{n}}^* \left(\boldsymbol{\mu}_{\boldsymbol{n}} \right) \right).$$

The primal formulation of the problem is thus linked to a dual formulation, which consists in maximizing the dual criterion $D(\boldsymbol{\mu}_{\boldsymbol{n}}) \triangleq \boldsymbol{\mu}_{\boldsymbol{n}}^t \boldsymbol{z} - \mathcal{F}_{\boldsymbol{x}}^* \left(\boldsymbol{H}^t \boldsymbol{\mu}_{\boldsymbol{n}} \right) - \mathcal{F}_{\boldsymbol{n}}^* \left(\boldsymbol{\mu}_{\boldsymbol{n}} \right)$. This amounts to selecting the parameters $\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \boldsymbol{H}^t \hat{\boldsymbol{\mu}}_{\boldsymbol{n}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{n}} \right)$ of the maximum entropy distribution belonging to the exponential family

$$\mathrm{d}p(\boldsymbol{u}, \boldsymbol{v}) = \exp \left(\boldsymbol{\mu}_{\boldsymbol{x}}^t \boldsymbol{u} + \boldsymbol{\mu}_{\boldsymbol{n}}^t \boldsymbol{v} - \mathcal{F}_{\boldsymbol{x}}^* \left(\boldsymbol{\mu}_{\boldsymbol{x}} \right) - \mathcal{F}_{\boldsymbol{n}}^* \left(\boldsymbol{\mu}_{\boldsymbol{n}} \right) \right) \mathrm{d}\nu_{\boldsymbol{x}}(\boldsymbol{u}) \ \mathrm{d}\nu_{\boldsymbol{n}}(\boldsymbol{v}).$$

From $\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}$, we derive $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{n}}$ using the so-called primal-dual relations:

$$\hat{\boldsymbol{x}} = \left. \frac{\partial \mathcal{F}_{\boldsymbol{x}}^* \left(\boldsymbol{\mu}_{\boldsymbol{x}} \right)}{\partial \boldsymbol{\mu}_{\boldsymbol{x}}} \right|_{\boldsymbol{\mu_x} = \hat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \boldsymbol{H}^t \hat{\boldsymbol{\mu}}_{\boldsymbol{n}}} \quad \text{and} \quad \hat{\boldsymbol{n}} = \left. \frac{\partial \mathcal{F}_{\boldsymbol{n}}^* \left(\boldsymbol{\mu}_{\boldsymbol{n}} \right)}{\partial \boldsymbol{\mu}_{\boldsymbol{n}}} \right|_{\boldsymbol{\mu_n} = \hat{\boldsymbol{\mu}}_{\boldsymbol{n}}}.$$

Remember also that $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{n}})$ is the mean of the maximum entropy distribution. By the way, we can notice that $\forall \boldsymbol{x}, \boldsymbol{n}, \boldsymbol{\mu_n} \quad \mathcal{F}(\boldsymbol{x}, \boldsymbol{n}) \geq D(\boldsymbol{\mu}_{\boldsymbol{n}})$ and also that $\mathcal{F}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{n}}) = D(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}})$ (dual attainment relation).

3.3. FURTHER DEVELOPMENTS

Both formulations of the method require analytical expression of the log-partition function $\mathcal{F}_{\boldsymbol{x}}^*$. It can't generally be computed if $\nu_{\boldsymbol{x}}$ isn't separable (well known exceptions to this rule are non separable Gaussian and exponential reference measures).

In the sequel, reasoning will thus be restricted to separable reference measures. In a previously exposed computation, we proved that separable reference measures yield separable primal cost functions. We introduce an additional vector $\boldsymbol{w}$ of reference measure $\nu_{\boldsymbol{w}}$ so as to penalize interactions between pixels. If we choose all measures to be separable, this yields a complete separable primal cost function

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{n}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{w}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{n}).$$

This cost function will be minimized subject to $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ and $\boldsymbol{w} = \boldsymbol{Q}\boldsymbol{x}$, $\boldsymbol{Q}$ being typically a differentiation matrix. We thus have:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_i \mathcal{F}_{\boldsymbol{x}_i}(x_i) + \sum_j \mathcal{F}_{\boldsymbol{w}_j} \left((\boldsymbol{Q}\boldsymbol{x})_j \right) + \sum_k \mathcal{F}_{\boldsymbol{n}_k} \left((\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})_k \right).$$

A straightforward computation similar to the one already detailed shows that the dual formulation of this problems writes:

$$\left\{ \begin{array}{rcl} D(\boldsymbol{\mu_w}, \boldsymbol{\mu_n}) & = & \boldsymbol{\mu}_{\boldsymbol{n}}^t \boldsymbol{z} - \mathcal{F}_{\boldsymbol{x}}^* \left(\boldsymbol{H}^t \boldsymbol{\mu}_{\boldsymbol{n}} - \boldsymbol{Q}^t \boldsymbol{\mu}_{\boldsymbol{w}} \right) - \mathcal{F}_{\boldsymbol{w}}^* \left(\boldsymbol{\mu}_{\boldsymbol{w}} \right) - \mathcal{F}_{\boldsymbol{n}}^* \left(\boldsymbol{\mu}_{\boldsymbol{n}} \right), \\ (\hat{\boldsymbol{\mu}}_{\boldsymbol{w}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{n}}) & = & \arg \max_{\boldsymbol{\mu_w}, \boldsymbol{\mu_n}} D(\boldsymbol{\mu_w}, \boldsymbol{\mu_n}), \\ \hat{\boldsymbol{\mu}}_{\boldsymbol{x}} & = & \boldsymbol{H}^t \hat{\boldsymbol{\mu}}_{\boldsymbol{n}} - \boldsymbol{Q}^t \hat{\boldsymbol{\mu}}_{\boldsymbol{w}}. \end{array} \right.$$

The MEM estimator is derived using

$$\hat{\boldsymbol{x}} = \left.\frac{\partial \mathcal{F}_{\boldsymbol{x}}^{*}(\boldsymbol{\mu}_{\boldsymbol{x}})}{\partial \boldsymbol{\mu}_{\boldsymbol{x}}}\right|_{\boldsymbol{\mu}_{\boldsymbol{x}}=\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}}, \quad \hat{\boldsymbol{w}} = \left.\frac{\partial \mathcal{F}_{\boldsymbol{w}}^{*}(\boldsymbol{\mu}_{\boldsymbol{w}})}{\partial \boldsymbol{\mu}_{\boldsymbol{w}}}\right|_{\boldsymbol{\mu}_{\boldsymbol{w}}=\hat{\boldsymbol{\mu}}_{\boldsymbol{w}}}, \quad \hat{\boldsymbol{n}} = \left.\frac{\partial \mathcal{F}_{\boldsymbol{n}}^{*}(\boldsymbol{\mu}_{\boldsymbol{n}})}{\partial \boldsymbol{\mu}_{\boldsymbol{n}}}\right|_{\boldsymbol{\mu}_{\boldsymbol{n}}=\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}}.$$

Selecting the dual variables $(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{w}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{n}})$ amounts to selecting the maximum entropy distribution in the exponential family

$$\begin{aligned} \mathrm{d}p(\boldsymbol{u},\boldsymbol{s},\boldsymbol{v}) &= \exp\left(\boldsymbol{\mu}_{\boldsymbol{x}}^{t}\boldsymbol{u} + \boldsymbol{\mu}_{\boldsymbol{w}}^{t}\boldsymbol{s} + \boldsymbol{\mu}_{\boldsymbol{n}}^{t}\boldsymbol{v} - \mathcal{F}_{\boldsymbol{x}}^{*}(\boldsymbol{\mu}_{\boldsymbol{x}}) - \mathcal{F}_{\boldsymbol{w}}^{*}(\boldsymbol{\mu}_{\boldsymbol{w}}) - \mathcal{F}_{\boldsymbol{n}}^{*}(\boldsymbol{\mu}_{\boldsymbol{n}})\right) \\ &\quad \mathrm{d}\nu_{\boldsymbol{x}}(\boldsymbol{u})\ \mathrm{d}\nu_{\boldsymbol{w}}(\boldsymbol{s})\ \mathrm{d}\nu_{\boldsymbol{n}}(\boldsymbol{v}). \end{aligned}$$

We chose the reference measure $\mathrm{d}\nu(\boldsymbol{u},\boldsymbol{s},\boldsymbol{w}) = \mathrm{d}\nu_{\boldsymbol{x}}(\boldsymbol{u})\ \mathrm{d}\nu_{\boldsymbol{w}}(\boldsymbol{s})\ \mathrm{d}\nu_{\boldsymbol{n}}(\boldsymbol{v})$ to be separable. This isn't a contradiction since the constraint $\boldsymbol{w} = \boldsymbol{Q}\boldsymbol{x}$ applies to the mean of random variables and not to the realizations of the random variables. This constraint indeed links the parameters of the maximum entropy distribution by the relation $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \boldsymbol{H}^{t}\hat{\boldsymbol{\mu}}_{\boldsymbol{n}} - \boldsymbol{Q}^{t}\hat{\boldsymbol{\mu}}_{\boldsymbol{w}}$.

In a Bayesian framework, it appears that minimizing $\mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{Q}\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})$ is equivalent to a MAP estimation scheme with prior distribution $p(\boldsymbol{x}) \propto \exp(-\mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) - \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{Q}\boldsymbol{x}))$ and likelihood $p(\boldsymbol{z}\,|\boldsymbol{x}) \propto \exp(-\mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}))$. The class of MEM estimators derived here is thus equivalent to the class of MAP estimators obtained with a given set of convex Markov random fields priors. Any MEM estimator is a MAP estimator and a MAP estimator is a MEM estimator providing there exists reference measures $\nu_{\boldsymbol{x}}, \nu_{\boldsymbol{w}}, \nu_{\boldsymbol{n}}$ verifying:

$$\begin{aligned} p(\boldsymbol{x}) &\propto \exp\left(-\mathrm{C}(\nu_{\boldsymbol{x}},\boldsymbol{x}) - \mathrm{C}(\nu_{\boldsymbol{w}},\boldsymbol{w})\right), \\ p(\boldsymbol{z}\,|\boldsymbol{x}) &\propto \exp\left(-\mathrm{C}(\nu_{\boldsymbol{n}},\boldsymbol{n})\right), \end{aligned}$$

where $\mathrm{C}(p,\boldsymbol{y})$ is the Cramér transform of the distribution p, the Cramér transform being applied to $\boldsymbol{y}$.

4. Simulations

We report two simulation examples. We first deal with a deconvolution example in spectroscopy and then with a Fourier synthesis example.

Three hyperparameters k_a, k_b, k_c are introduced, the cost function to be minimized being $\mathcal{F}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{n}) = k_a\mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + k_b\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{w}) + k_c\mathcal{F}_{\boldsymbol{n}}(\boldsymbol{n})$, subject to $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ and $\boldsymbol{w} = \boldsymbol{Q}\boldsymbol{x}$. The corresponding dual function to be maximized is $\mathcal{F}^{*}(\boldsymbol{\mu}_{\boldsymbol{x}},\boldsymbol{\mu}_{\boldsymbol{w}},\boldsymbol{\mu}_{\boldsymbol{n}}) = k_c\boldsymbol{\mu}_{\boldsymbol{n}}^{t}\boldsymbol{z} - k_a\mathcal{F}_{\boldsymbol{x}}^{*}(\boldsymbol{\mu}_{\boldsymbol{x}}) - k_b\mathcal{F}_{\boldsymbol{w}}^{*}(\boldsymbol{\mu}_{\boldsymbol{w}}) - k_c\mathcal{F}_{\boldsymbol{n}}^{*}(\boldsymbol{\mu}_{\boldsymbol{n}})$ subject to $\boldsymbol{\mu}_{\boldsymbol{x}} = \left(k_c\boldsymbol{H}^{t}\boldsymbol{\mu}_{\boldsymbol{n}} - k_b\boldsymbol{Q}^{t}\boldsymbol{\mu}_{\boldsymbol{w}}\right)/k_a$. The hyperparameters are chosen empirically.

We tried several optimization schemes out (primal and dual objective functions optimized with pseudo–conjugate gradient descent method and pixel by pixel (Gauss–Seidel) descent method). It seems hard to derive a definite general rule from the simulations we carried out. In any case, it must be checked if optimization has been properly completed. In the primal domain, one will thus have to check if

$$\mathcal{F}(\hat{\boldsymbol{x}},\hat{\boldsymbol{w}},\hat{\boldsymbol{n}}) = \mathcal{F}^{*}\left(\left.\frac{\partial \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}, \left.\frac{\partial \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{w})}{\partial \boldsymbol{w}}\right|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}, \left.\frac{\partial \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{n})}{\partial \boldsymbol{n}}\right|_{\boldsymbol{n}=\hat{\boldsymbol{n}}}\right).$$

A similar relation has to be checked in the dual domain.

Moreover, one should be careful not to stop iterations before the optimum is reached when operating in the dual variables. This would yield unsatisfactory results, much more unsatisfactory than when stopping iterations before the optimum is reached in the primal variables.

Simulations we carried out speak in favor of pixel by pixel minimization. On the one hand, given a pixel to be updated, every function evaluation is achieved at a very low computational cost (instead of computing $\boldsymbol{H}\boldsymbol{x}$ or $\boldsymbol{H}^t\boldsymbol{\mu_n}$, we use updating formulas). Gradient or pseudo–conjugate gradient descent requires computation of $\boldsymbol{H}\boldsymbol{x}$ or $\boldsymbol{H}^t\boldsymbol{\mu_n}$ for any function evaluation. On the other hand, separation of the variables in the pixel by pixel scheme allows more efficient optimization of the function (figure $2c$).

4.1. DECONVOLUTION

We used a mixture of two Gamma laws as a reference measure for the pixels and a Gaussian law as a reference measure for the noise.

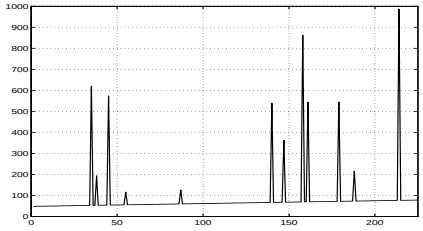

Fig. 1a: original signal

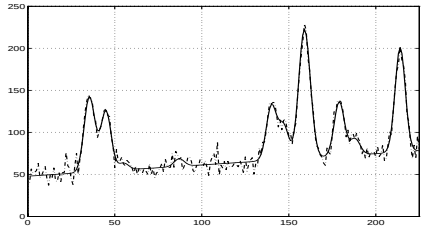

Fig. 1b: observations; SNR : 16 dB

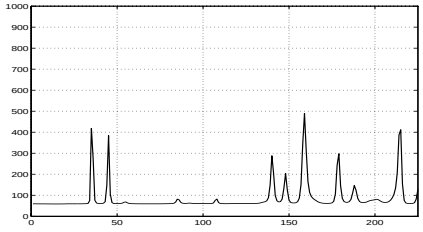

Fig. 1c: mem restauration

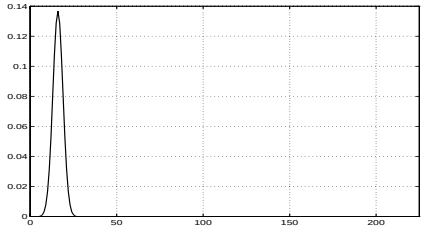

Fig. 1d: convolution kernel

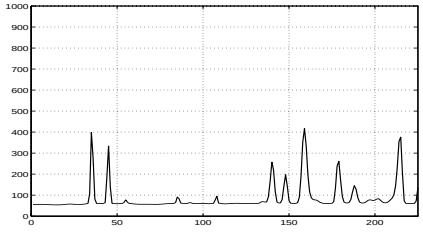

Fig. 1e: Lp restauration (p=1.1)

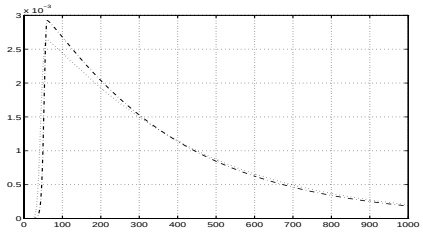

Fig. 1f: priors

Figure $1f$ (priors) depicts $\exp(-x^p)$ (dotted line) and $\exp(-\mathcal{F}(x))$ (dashdotted line). Those functions would play the role of a prior distribution in a Bayesian

framework.

4.2. FOURIER SYNTHESIS

This example is taken from [11]. The data are the real parts of the twenty first Fourier coefficients which are supposed to be perfectly known. We used a mixture of two exponential laws (defined on R^+) as a reference measure for the pixels, a mixture of two symmetric exponential laws as a reference measure for the difference between neighboring pixels and a Gaussian law as a reference measure for the noise.

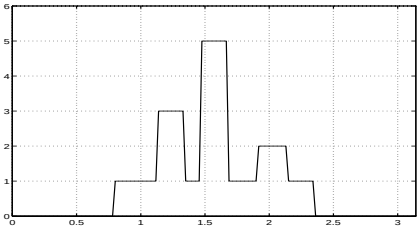

Fig. 2a: original signal

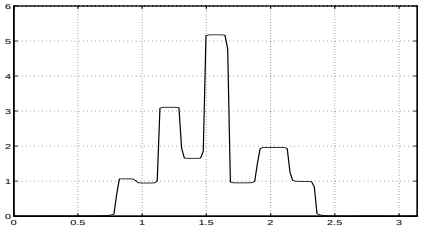

Fig. 2b: mem reconstruction

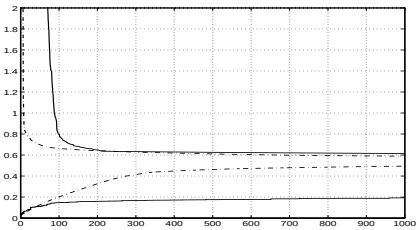

Fig. 2c: criteria *vs* cpu time (seconds)

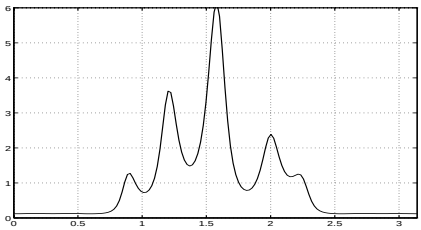

Fig. 2d: reconstruction; seperable criterion

Figure 2*c* depicts the value of the objective functions (above: primal domain; below: dual domain; solid lines: pseudo–conjugate gradient descent; dashdotted lines: pixel by pixel descent). Figure 2*d* depicts reconstruction obtained without penalizing differences between pixels, which would be comparable to the results reported in [11].

5. Conclusion

This communication deals with both theoretical and algorithmic issues. On the one hand, we account for penalization of interactions between pixels in the MEM formalism. This introduces additional terms in the criterion, which nevertheless remains strictly convex. On the other hand, we focus on implementation issues. This speaks in favor of pixel by pixel minimization (Gauss–Seidel strategy), either in the primal or in the dual domain. It appears in the case investigated here that minimization in the primal domain is less expensive than in the dual domain. This doesn't hold as a general and definite conclusion. It merely shows that many strategies are available and that the choice of the suitable one is certainly problem dependant.

Additional work on the method is under way. Other approaches for penalizing interactions in the MEMM have been proposed: Urban [12] decomposes the error n_i as $n_i = n_i^0 + \sum_j n_{ij}$, where $n_{ij} = n_{ji}$ and uses a separable reference measure on $\{n_i^0, n_{ij}\}$. Moreover, study of mathematical aspects of the method by Csiszár, Gamboa and Gassiat is in progress.

The MEM provides a framework for constructing convex criteria. This framework also allows to interpret well known criteria. But, in the opinion of the authors, this framework is by no mean exclusive. In particular, [2] suggests many interesting convex and nonconvex criteria in a Bayesian approach.

References

1. J.-F. Bercher, G. Le Besnerais, and G. Demoment, *The maximum entropy on the mean method, noise and sensitivity.* Maximum entropy and Bayesian methods, Kluwer Academic Publishers, Cambridge, U.K., 1995.
2. C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Transactions on Image Processing*, **IP-2**, pp. 296–310, July 1993.
3. F. Gamboa, *Méthode du maximum d'entropie sur la moyenne et applications.* PhD thesis, Université de Paris-Sud, Orsay, December 1989.
4. J. Navaza, "On the maximum entropy estimate of electron density function," *Acta Crystallographica*, **A–41**, pp. 232–244, 1985.
5. J. Navaza, "The use of non-local constraints in maximum-entropy electron density reconstruction," *Acta Crystallographica*, **A–42**, pp. 212–223, 1986.
6. D. Dacunha-Castelle and F. Gamboa, "Maximum d'entropie et problème des moments," *Annales de l'Institut Henri Poincaré*, **26**, (4), pp. 567–596, 1990.
7. G. Le Besnerais, *Méthode du maximum d'entropie sur la moyenne, critères de reconstruction d'image, et synthèse d'ouverture en radio-astronomie.* PhD thesis, Université de Paris-Sud, Orsay, December 1993.
8. J.-F. Bercher, *Développement de critères de nature entropique pour la résolution des problèmes inverses linéaires.* PhD thesis, Université de Paris-Sud, Orsay, February 1995.
9. R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*, Springer-Verlag, New York, 1985.
10. C. Heinrich, J.-F. Bercher, G. Le Besnerais, and G. Demoment, "Restoration of spiky signals: a new optimal estimate and a comparison," in *Proceedings of IEEE ICASSP*, (Detroit, U.S.A.), pp. 877–880, May 1995.
11. L. K. Jones and V. Trutzer, "Computationally feasible high-resolution minimum-distance procedures which extend the maximum-entropy method," *Inverse Problems*, **5**, pp. 749–766, 1989.
12. B. Urban, "Retrieval of atmospheric thermodynamical parameters using satellite measurements with a maximum entropy method," tech. rep., Centre National de Recherches Météorologiques, Toulouse, France, April 1996.